

## 明細書

## 対象音検出方法、信号入力遅延時間検出方法及び音信号処理装置

## 技術分野

- 5 本発明は、検出対象音を検出する対象音検出方法及びそのプログラム、複数のマイクロホンに入力される音信号間の遅延時間を検出する信号入力遅延時間検出方法及びそのプログラム、入力された音信号を処理する音信号処理装置、並びに発話音を検出し、その発話音について音声認識処理を行う音声認識装置に関する。

10

## 背景技術

- 音声は、人間の用いる種々の通信の形態の中でも最も根源的であると同時に、他のどの情報送出方法よりも高速度に情報を送り出すことのできる優れた通信手段である。このようなことから、音声は、古くから現在に至るまで人間の通信手段の根幹を担ってきた。

- また、そのような音声を認識するための音声認識技術がある。音声認識とは、その音声に含まれる情報の中で、最も基本的な意味内容に関する情報、つまり音韻情報をコンピュータなどにより抽出し、その抽出内容を判定することである。近年では、計算機プロセッサ技術の飛躍的な発達と、インターネットに代表される高度な情報ネットワークの構築により、様々な分野においてマン・マシンインタフェースとしての音声認識技術の適用が試みられている。

- 現在の音声認識システムの認識性能は、確率・統計的手法により格段に向上しており、理想的な環境下での音声や接話マイクロホンで収録された近距離音声などでは、非常に高い認識率が得られるようになっている。

- 25 ところで、実環境下の音声認識は、学習データと観測データとの間の環境、発話内容などのミスマッチにより、その認識率が劣化する。また、受音系となる接話マイクヘッドセットの装着によりユーザが受ける負担や不快感は大きく、音声認識システム実用化の大きな障害のひとつになっている。

また、S/N比の低下や背景雑音、室内残響の影響などにより認識が困難な遠

隔音声に関し、複数の遠隔マイクロホンを用いた音声認識手法の研究が多くなされている。その代表的なものとして、マイクロホンアレーを用いる手法が挙げられる。この手法では、音源位置検出処理、目的音強調処理、雑音抑制処理、の3つの空間的な信号処理を行なうことができる。このような手法により遠隔音声の

5 音声認識が盛んに研究されている。

しかし、この手法は、正確な話者方向同定処理のために複数のマイクロホンを一定間隔にて固定配置する必要があり、小型化、携帯化が困難であるため、様々な環境・状況下での音声入力への応用が難しく、用途が限定されるという問題がある。

- 10 ここで、いつでもどこでも音声入力を可能にするユビキタスな受音系として、①小型・軽量で脱着が容易、②接話マイクとほぼ同等の近距離音声を確認することができる、③接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる、という点で、衣服や眼鏡などに取り付けることができる装着型マイクロホンが期待されている。

- 15 本発明は、前述の問題に鑑みてなされたものであり、複数の装着型マイクロホンを用いた環境変動に対してもロバストな受音系の構築を可能にする対象音検出方法、信号入力遅延時間検出方法、音信号処理装置、音声認識装置及プログラムの提供を目的とする。

20 発明の開示

- 本発明に係る対象音検出方法は、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相  
25 の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、前記帯域毎のそれぞれの傾きが特定の傾

きに集中する傾向が強くなったときに検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする。

- 5      また、本発明に係る信号入力遅延時間検出方法は、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出することを特徴とする。
- 10

また、前記信号入力遅延時間検出方法において、前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記受音の遅延時間を検出することを特徴とする。

- また、前記信号入力遅延時間検出方法において、前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに、前記受音の遅延時間を検出することを特徴とする。
- 15

また、前記信号入力遅延時間検出方法において、複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする。

- 20      また、本発明に係る音信号処理装置は、複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した検出対象音源から出力された検出対象音を検出する対象音検出手段と、を備えたことを特徴とする
- 25

また、前記音信号処理装置において、前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、

前記検出対象音を検出することを特徴とする。

- また、本発明に係る音信号処理装置は、音源から出力された音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された音进行处理する音信号処理装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、を備えたことを特徴とする。

また、前記音信号処理装置において、前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、

- また、前記音信号処理装置において、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出することを特徴とする。

- また、本発明に係る音信号処理装置は、検出対象音源から出力された検出対象音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された検出対象音进行处理する音信号処理装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記検出対象音源からの受音の遅延時間を検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記検出対象音を検出する対象音検出手段と、を備えたことを特徴とする。

また、前記音信号処理装置において、前記傾き検出手段は、前記クロススペクト

ルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記検出対象音を検出することを特徴とする。

- 5      また、本発明に係る音声認識装置は、発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音进行处理する音声認識装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを  
10    検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した前記発話音を検出する発話音検出手段と、前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、を備えたことを特徴とする。

- また、前記音声認識装置において、前記傾き検出手段は、前記クロススペクトル  
15    ルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする。

- また、本発明に係る音声認識装置は、発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音进行处理する  
20    音声認識装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを  
検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾き  
に基づいて、前記複数のマイクロホン間での前記発話源からの受音の遅延時間を  
25    検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記発話音を検出する発話音検出手段と、前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声

認識処理手段と、を備えたことを特徴とする。

また、前記音声認識装置において、前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音  
5 の遅延時間を検出し、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする。

また、本発明に係るプログラムは、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイ  
10 クロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検出対象音を検出する処理をコンピュータに実行させることを特徴とする。

また、本発明に係るプログラムは、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロス  
15 スペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させることを特徴とする。

ここで、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源と各マイクロホンとの間のそれぞれの距離の差に対応して、その位相の周波数に対する傾きが一定になる。そして、音源と各マイクロホンとの間のそれぞれの距離の差は、複数のマイクロホン間での受音の遅延時間として現れる。さらに、複数のマイクロホンで受音した音声のS/N比が高ければ、そ  
25 のように傾きが一定となる傾向が顕著になる。本発明はこのような関係を利用したものである。

すなわち、本発明では、複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出

し、その傾きに基づいて、当該複数のマイクロホンが受音した検出対象音や発話音を検出している。なお、検出対象音には、人間が発する発話音の他、物体が発する音も含まれる。

この発明は、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源から各マイクロホンとの距離の差に対応して、その位相の周波数に対する傾きが一定になり、その一方で、複数のマイクロホンで受音した音のS/Nが高ければ、そのように傾きが一定となる傾向が顕著になること、を原理としたものである。

また、本発明では、複数の複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、音源と前記複数のマイクロホンとの間のそれぞれの距離の差に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での受音の遅延時間を検出している。

この発明は、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源と各マイクロホンとの間のそれぞれの距離の差に対応して、その位相の周波数に対する傾きが一定になり、その一方で、音源と各マイクロホンとの間のそれぞれの距離の差が、複数のマイクロホン間での受音の遅延時間として現れること、を原理とするものである。

また、本発明では、クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎の前記傾きに基づいて処理を行っている。これにより、精度を上げて前記傾きを検出している。

#### 図面の簡単な説明

図1は、本発明の実施の形態の音声信号処理装置を含むシステム全体の構成を示すブロック図である。図2は、本発明の第1の実施の形態の音声信号処理装置の構成を示すブロック図である。図3は、各環境のクロススペクトルの位相を示す特性図である。図4は、クロススペクトルの位相を示す特性図であり、(A)は、音声区間フレームのクロススペクトルの位相を示す特性図であり、(B)は、非音声区間フレームのクロススペクトルの位相を示す特性図である。図5は、

クロススペクトルの位相に基づいて得たヒストグラムを示す特性図であり、(A)は、音声区間フレームのヒストグラムを示す特性図であり、(B)は、非音声区間フレームのヒストグラムを示す特性図である。図6は、音声信号処理装置のヒストグラム等計算部などの構成を示すブロック図である。図7は、第1の実施  
5 形態の音声信号処理装置の効果の説明に用いた特性図である。図8は、本発明の第2の実施の形態の音声信号処理装置の構成を示すブロック図である。図9は、合成信号を生成するためのオーバーラップアッド法の説明に用いた図である。図10は、第2の実施の形態の音声信号処理装置の効果の説明に用いた特性図である。図11は、本発明の第3の実施の形態の音声信号処理装置の構成を示すブ  
10 ロック図である。図12は、音声信号処理装置の音声／非音声判定部の他の構成を示すブロック図である。

#### 発明を実施するための最良の形態

以下、本発明の実施の形態を図面を参照しながら詳細に説明する。この実施の  
15 形態は、図1に示すように、2つのマイク1、2で受音した音声信号を処理する音声信号処理装置10である。ここで、第1及び第2マイク1、2は音源（ユーザ）自体に比較的自由度を持たせた位置に装着可能な装着型マイクである。

図2は、第1の実施の形態の音声信号処理装置10の構成を示す。図2に示すように、音声信号処理装置10は、第1及び第2フレーム化部11、12と、第  
20 1及び第2周波数分析部13、14と、クロススペクトル計算部15と、位相抽出処理部16と、位相 *unwrap* 処理部17と、主計算部30と、音入力オン／オフ制御部18とを備えている。また、主計算部30については、周波数帯域分割部31と、第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>と、ヒストグラム等計算部33と、音声／非音声判定部34とを備えている。以下、各部の処理内容を説明する。

25 第1及び第2マイク1、2から入力された2chの音声信号はそれぞれ、第1及び第2フレーム化部11、12に入力される。また、第1マイク1から入力された音声信号は、音入力オン／オフ制御部18に入力される。



第1及び第2フレーム化部11, 12、第1及び第2周波数分析部13, 14及びクロススペクトル計算部15により、第1及び第2マイク1, 2から入力された2chの音声信号のクロススペクトルを算出する。

例えば、第1マイク1と第2マイク2といった複数のマイクで受音した音声信号を時間軸上でみた場合、受音した音声信号間に位相差が生じる。これは、音源から各マイク1, 2までの距離の違いにより、音源から各マイク1, 2までの音声信号の到達時間に差が生じた結果である。

ここで、第1マイク1と第2マイク2とにより受音した音声信号間の遅延時間を計測し、その計測した遅延時間に基づいて位相を同相化し、その後、第1マイク1と第2マイクとでそれぞれ受音した音声信号を加算して同期加算音声を得る場合を考える。例えば、M.Omologo, P.Svaizer らの文献「"Acoustic event localization using a crosspower-spectrum phase based technique", Proc.ICASSP94, pp.274-276, (1994)」に、そのように同期加算音声を得る技術が記載されている。

ここで、2つのマイク1, 2で受音した音声信号をそれぞれ $x_1(t)$ ,  $x_2(t)$ とし、これら $x_1(t)$ ,  $x_2(t)$ をフーリエ変換して得られる周波数関数を $X_1(\omega)$ ,  $X_2(\omega)$ とする。ここで、 $x_2(t)$ は、下記(1)式のように $x_1(t)$ の時間移動波形であると仮定する。

$$x_2(t) = x_1(t - t_0) \quad \dots (1)$$

このように仮定した場合、周波数関数 $X_1(\omega)$ と周波数関数 $X_2(\omega)$ との関係は下記(2)式のようになる。

$$X_2(\omega) = e^{-j\omega t_0} X_1(\omega) \quad \dots (2)$$

そして、この周波数関数 $X_1(\omega)$ と周波数関数 $X_2(\omega)$ とからクロススペクトル $G_{12}(\omega)$ が下記(3)式として得られる。

$$G_{12}(\omega) = X_1(\omega) X_2^*(\omega) = X_1(\omega) e^{j\omega t_0} X_1^*(\omega) = |X_1|^2 e^{j\omega t_0} \quad \dots (3)$$

ここで、クロススペクトル $G_{12}(\omega)$ の指数項はスペクトル領域のチャネル間の時間遅れに対応する。したがって、周波数関数 $X_2$ に遅延項 $e^{j\omega t_0}$ をかけた $X_2(\omega) e^{j\omega t_0}$ は、周波数関数 $X_1$ と同相化され、これにより、 $X_1(\omega) + X_2$

( $\omega$ )  $e^{j\omega t}$  の逆フーリエ変換をチャンネル同期加算音声として扱うことができるようになる。

クロススペクトル計算部 15 により、このようなクロススペクトル  $G_{12}(\omega)$  を得る。

- 5     そのため、まず、第 1 フレーム化部 11 では、後段の第 1 周波数分析部 13 のために、第 1 マイク 1 から入力される音声信号をフレーム化（或いはフレーム分割）して、当該第 1 周波数分析部 13 に出力する。また、第 2 フレーム化部 12 では、後段の第 2 周波数分析部 14 のために、第 2 マイク 2 から入力される音声信号をフレーム化（或いはフレーム分割）して、当該第 2 周波数分析部 14 に出力する。第 1 及び第 2 フレーム化部 11、12 は、所定のサンプル数を 1 フレームとして、入力されてくる音声信号を次々にフレーム化していく。
- 10

例えば、マイク 1、2 に音声が入力（発話入力）されていない場合には、フレームは、音声の入力されていない非音声区間フレームとなり、マイク 1、2 に音声が入力されている場合には、フレームは、音声の入力（発話入力）されている

- 15   音声区間フレームとなる。

- 第 1 周波数分析部 13 は、第 1 フレーム化部 11 からの音声信号をフーリエ変換して周波数関数  $X_1(\omega)$  を算出して、後段のクロススペクトル計算部 15 に出力する。また、第 2 周波数分析部 14 は、第 2 フレーム化部 12 からの音声信号をフーリエ変換して周波数関数  $X_2(\omega)$  を算出して、後段のクロススペクトル計算部 15 に出力する。ここで、第 1 及び第 2 周波数分析部 13、14 は、フレーム毎に音声信号をフーリエ変換する。
- 20

クロススペクトル計算部 15 は、第 1 及び第 2 周波数分析部 13、14 からの周波数関数  $X_1(\omega)$ 、 $X_2(\omega)$  に基づいて、前記 (3) 式によりクロススペクトル  $G_{12}(\omega)$  を算出する。

- 25   なお、図 3 には、1 フレームについての音声信号のクロススペクトルの位相を示しており、図 3 中 (A) は自動車内で発した音声について得たクロススペクトルの位相であり、図 3 中 (B) はオフィススペース内で発した音声について得たクロススペクトルの位相であり、図 3 中 (C) は防音室内で発した音声について得たクロススペクトルの位相であり、図 3 中 (D) は歩道（屋外）で発した音声

について得たクロススペクトルの位相である。この図3に示すように、フレーム内で、すなわち局所的に、音源と第1マイク1までの距離と音源と第2マイク2までの距離との差に対応して、クロススペクトルの位相が周波数に対してほぼ一定の傾きを示すことがわかる。すなわち、音源と第1マイク1までの距離と音源と第2マイク2までの距離との差に対応して、クロススペクトルの位相成分が一定の傾きを有している。

また、第1及び第2マイク1, 2で受信した音声信号のS/N比が高ければ、そのように傾きが一定となる傾向は顕著になるのである。ここで、第1及び第2マイク1, 2が装着型マイクなので、第1及び第2マイク1, 2により音声を受音した場合のその音声信号はS/N比が高くなり、このようなことから、明らかに一定の傾きを示すものになっている。

クロススペクトル計算部15は、このような特性を有するクロススペクトル $G_{12}(\omega)$ を位相抽出部16に出力する。

位相抽出部16では、クロススペクトル計算部15からのクロススペクトル $G_{12}(\omega)$ から位相を抽出（検出）して、その抽出結果を位相unwrap処理部17に出力する。

位相unwrap処理部17では、位相抽出部16の位相抽出結果に基づいて、クロススペクトル $G_{12}(\omega)$ をunwrap処理して、主計算部30の周波数帯域分割部31に出力する。

周波数帯域分割部31は、帯域分割（セグメント分割）した位相を第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>それぞれに出力する。

ここで、音声の入力されていない非音声区間フレームと音声が入力されている音声区間フレームとで、クロススペクトルの位相成分に大きな違いがある。すなわち、音声区間フレームでは、前述したようにクロススペクトルの位相が周波数に対してほぼ一定の傾きを示すが、非音声区間フレームでは、そのようにはならない。ここで、図4を用いて説明する。

図4はクロススペクトル（CRS）の位相を示しており、図4中（A）は、音声区間フレームのクロススペクトルの位相であり、図4中（B）は、非音声区間フレームのクロススペクトルの位相である。

この図4中(A)と図4中(B)との比較からもわかるように、非音声区間フレームでは、クロススペクトルの位相は、周波数に対して特定のトレンドをもたないものである。すなわち、周波数に対してクロススペクトルの位相が一定の傾きを持つ結果とはならない。これは、ノイズの位相がランダムだからである。

- 5      これに対して、音声区間フレームでは、周波数に対してクロススペクトルの位相が一定の傾きをもつようになる。そして、この傾きは、音源から各マイク1, 2までの距離の差に対応した大きさになる。

このように、音声の入力されていない非音声区間フレームと音声が入力されている音声区間フレームとでは、クロススペクトルの位相成分に大きな違いがある。

- 10     このようなことから、位相の回転が生じた場合にも正確にトレンドを追従するために、周波数帯域分割部31により、位相成分を小さな周波数セグメントに分割（或いは帯域分割）し、後段の第1乃至第N傾き計算部32<sub>1</sub>~32<sub>N</sub>で、最小2乗法を適用することでセグメント毎に傾きを計算している。この第1乃至第N傾き計算部32<sub>1</sub>~32<sub>N</sub>はそれぞれ、算出した傾きをヒストグラム等計算部33  
15     に出力する。

ここで、最小2乗法によりセグメント毎に傾きを求める手法は、公知の技術であり、例えば、『「信号処理」「画像処理」のための入門工学社、高井信勝著、工学社、(2000)』にその技術が記載されている。

- 20     ヒストグラム等計算部33は、第1乃至第N傾き計算部32<sub>1</sub>~32<sub>N</sub>が算出した前記傾きについて、ヒストグラムを得る。

- 図5は、ヒストグラム等計算部33が得たヒストグラムで、セグメント毎に得た傾きについてのヒストグラムを示している。すなわち、この図5は、位相の傾きの分布を示し、全セグメントに対する、各傾きのセグメント数の割合、すなわち頻度を縦軸にとっている。ここで、図5中(A)は、音声区間フレームについて  
25     のヒストグラムを示し、図5中(B)は、非音声区間フレームについてのヒストグラムを示す。

この図5中(A)と図5中(B)との比較からもわかるように、音声区間フレームでは、ヒストグラムに明らかにピーク値があり、すなわち傾きがごく狭い範囲に局在して、これにより、ある傾きについて頻度が高くなっている。すなわち、

帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなっている。一方、非音声区間フレームでは、ヒストグラムが平滑となり、傾きが広い範囲にわたって分布している。

- このヒストグラム等計算部 33 は、このようなヒストグラム化して得た頻度を  
5 音声／非音声判定部 34 に出力する。なお、このヒストグラム等計算部 33 の処理については後で具体例を説明する。

- 音声／非音声判定部 34 は、ヒストグラム等計算部 33 からの前記頻度に基づいて、音声区間と非音声区間とを判定する。例えば、前記頻度の平均値周辺の所定の範囲に含まれる傾きの出現頻度が所定の閾値以上の場合、音声区間と判定し、  
10 頻度が所定の閾値未満の場合、非音声区間と判定する。

なお、ここでは、前段の処理がフレーム単位の処理となっているので、当該フレームが、音声区間フレーム又は非音声区間フレームのいずれかであるかを判定する。音声／非音声判定部 34 は、その判定結果を音入力オン／オフ制御部 18 に出力する。

- 15 音入力オン／オフ制御部 18 には、第 1 マイク 1 からの音声信号が入力されており、音入力オン／オフ制御部 18 は、音声／非音声判定部 34 の判定結果に基づいて、その第 1 マイク 1 からの音声信号の後段への出力をオンとオフとを切り換える。具体的には、音声／非音声判定部 34 が音声区間と判定した場合、音入力オン／オフ制御部 18 は、オンにして音声信号を後段に出力して、音声／非  
20 音声判定部 34 が非音声区間と判定した場合、音入力オン／オフ制御部 18 は、オフにして音声信号を後段に出力しないようにする。

なお、前段の処理がフレーム単位の処理となっているので、音入力オン／オフ制御部 18 は、判定対象のフレームに対応した第 1 マイク 1 からの音声信号の部位を単位としてオンとオフとを切り換える。

- 25 ヒストグラム等計算部 33 の処理の具体例を説明する。図 6 は、その処理を実現するヒストグラム等計算部 33 の構成を示す。

ヒストグラム等計算部 33 は、第 1 乃至第 N 傾き計算部 32<sub>1</sub>～32<sub>N</sub>が算出した前記傾きのうちから頻度が高い（最頻度の）傾きを算出する構成として、第 1 スイッチ 33S1、第 2 スイッチ 33S2 及び最頻値計算部 33C を備えている。

これにより、第1スイッチ33S1を一定時間オン（閉）にして、第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>が算出した一定時間の前記傾きのデータ（或いはデータベース）33D1を作成する。このとき、第2スイッチ33S2については、オフ（開）にしておく。そして、データ33D1を作成したら、第2スイッチ33S2をオン（閉）にして、そのデータ33D1を最頻値計算部33Cに出力する。

最頻値計算部33Cでは、データ33D1から前記図5に示すような前記傾きについてのヒストグラムを作成して、そのヒストグラム中の最頻度の傾き（以下、最頻傾きという。） $\tau_0$ を算出する。なお、最頻度の傾きを算出するようにしてもよいが、平均値の傾き $\tau_0$ を算出したり、或いは最頻度の傾きと傾きの平均値とを組み合わせた傾き $\tau_0$ を算出するようにしてもよい。これにより、各帯域の傾きが特定の傾きに集中する傾向が強くなったとき、当該特定の傾きの値そのもの或いはそれに近い傾きの値を得ることができる。なお、本実施の形態では、最頻値計算部33Cが最頻傾き $\tau_0$ を算出しているものとする。

そして、最頻値計算部33Cは、算出した最頻傾き $\tau_0$ を前記音声／非音声判定部34に出力する。ここで、最頻傾き $\tau_0$ をデータ33D2として前記音声／非音声判定部34に出力する。

以上がヒストグラム等計算部33の処理の具体例である。

前記音声／非音声判定部34では、ヒストグラム等計算部33からの最頻傾き $\tau_0$ に基づいて、音声区間と非音声区間とを判定する。

なお、先の説明では、音声／非音声判定部34がヒストグラム等計算部33からの前記傾度に基づいて音声区間と非音声区間とを判定する場合について説明した。ここでは、音声／非音声判定部34は、ヒストグラム等計算部33からの最頻傾き $\tau_0$ と第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>が算出した前記傾き（各帯域の傾き） $\tau_i$ に基づいて、音声区間と非音声区間とを判定しており、これに対応して、音声／非音声判定部34に、第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>が算出した前記傾きが入力されるようになっている。

すなわち、音声／非音声判定部34は、第1乃至第N傾き計算部32<sub>1</sub>～32<sub>N</sub>が算出した前記傾き $\tau_i$ と最頻傾き $\tau_0$ とを下記（4）式により比較する。

$$|\tau_i - \tau_0| < \delta \quad \dots (4)$$

ここで、 $\delta$ は判定用の閾値（傾き閾値）である。

音声／非音声判定部 34は、この（4）式の条件が満たされていることが所定の割合を超えた場合（YES）、音声区間と判定し、そうでない場合（NO）、非音声区間と判定する。そして、音声／非音声判定部 34は、その判定結果を音入力オン／オフ制御部 18に出力する。

以上のように構成した音声信号処理装置 10の一連の動作は次のようになる。

5      先ず、第1及び第2フレーム化部 11、12、第1及び第2周波数分析部 13、14及びクロススペクトル計算部 15が、第1及び第2マイク 1、2から入力された2chの音声信号のクロススペクトル  $G_{12}(\omega)$  を算出する。

10      そして、位相抽出部 16、位相 unwrap 処理部 17及び周波数帯域分割部 31が、そのように算出したクロススペクトル  $G_{12}(\omega)$  の位相を帯域分割（セグメント分割）して、第1乃至第N傾き計算部 32<sub>1</sub>～32<sub>N</sub>が、帯域毎（セグメント毎）の位相の傾きを算出する。

そして、ヒストグラム等計算部 33が、第1乃至第N傾き計算部 32<sub>1</sub>～32<sub>N</sub>

15      それぞれが算出した前記帯域毎（セグメント毎）の傾きからヒストグラムを生成して、音声／非音声判定部 34が、そのヒストグラムから得られる頻度と最頻傾き  $\tau_0$  に基づいて、音声区間と非音声区間とを判定する。この判定結果に基づいて、音入力オン／オフ制御部 18では、第1マイク 1からの音声信号の後段への出力をオンとオフとを切り換える。具体的には、音声／非音声判定部 34が音声

20      区間と判定した場合、音入力オン／オフ制御部 18は、オンにして音声信号を後段に出力して、音声／非音声判定部 34が非音声区間と判定した場合、音入力オン／オフ制御部 18は、オフにして音声信号を後段に出力しないようにする。

このように、音声信号処理装置 10は、第1マイク 1、2が受音した音声中の発話区間（音声区間）を検出することができる。

25      例えば、第1マイク 1、2と音声アプリケーションとの間にこのような音声信号処理装置 10を備えることで、音声アプリケーションは、確実に発話区間についての処理を行うことができる。ここで、音声アプリケーションとしては、音声認識システム、放送システム、携帯電話及びトランシーバが挙げられる。例えば、音声アプリケーションが音声認識システムであるとすれば、音声認識システムは、

音声信号処理装置 10 が出力する発話区間の音声信号に基づいて音声認識することができるようになる。

次に効果を説明する。

- 前述したように、第 1 及び第 2 マイク 1, 2 に入力された音信号間のクロス  
5    スペクトルの位相を検出し、その検出したクロススペクトルの位相の周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した音声信号中の発話区間を検出している。すなわち、音声が入力（発話入力）されていない音声信号と音声が入力（発話入力）されている音声信号とをクロススペクトルでみた場合に、そのクロススペクトルの位相成分に大きな違いがあることを利用して、当該複数の  
10    マイクロホンが受音した音声信号中の発話区間を検出している。

具体的には、クロススペクトルの位相を帯域分割（セグメント分割）し、帯域  
毎（セグメント毎）の位相の傾きからヒストグラムを生成し、そのヒストグラムから頻度（具体的には最頻値）を得て、その頻度に基づいて、発話区間を検出している。

- 15    これにより、精度よく発話区間を検出することができる。そして、このように音声信号処理装置 10 が検出した発話区間の音声信号を利用することにより、音声認識システムでは、高認識率、低誤認識率の音声認識が可能になり、また、携帯電話やトランシーバでは、信頼性の高いハンズフリー半二重通信が可能になり、放送システムでは、通信システムの送信電力低減が可能になる。

- 20    また、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。

- 前述したように、クロススペクトルの位相の周波数に対する傾きは、音源と第  
1 マイク 1 までの距離と音源と第 2 マイク 2 までの距離との差に対応して変化する値になっている。これにより、例えば、音源に対する第 1 及び第 2 マイク 1,  
25    2 の取り付け位置を変更した場合、クロススペクトルの位相の周波数に対する傾きはその位置の変更に対応して変化するようになる。その一方で、前述したように、クロススペクトルの位相を帯域分割（セグメント分割）し、帯域毎（セグメント毎）の位相の傾きからヒストグラムを生成し、そのヒストグラムから頻度（具体的には最頻度）を得て、その頻度に基づいて、発話区間を検出している。



すなわち、クロススペクトルの位相の傾きの大きさ自体に拠ることなく、つまり、音源とマイク 1, 2 との間の距離に左右されることなく、最終的に、発話区間の検出を行っている。よって、音源に対する第 1 及び第 2 マイク 1, 2 の取り付け位置を変更した場合でも、発話区間の検出結果への影響はない。

- 5      このようなことから、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

- 10      以上のように、小型・軽量で脱着が容易であり、接話マイクとほぼ同等の近距離音声を確認することができ、接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる装着型マイクを用いることを前提としつつも、前述した種々の効果を得ることができる。

(実施例 (第 1 の実施の形態))

- 20      本発明を適用したシステムにより音声の発話区間の検出を行った。各文章間に 1 秒程度の無発話区間を含む合計 40 文をサンプルの使用音声とした。実験環境は、防音室内、自動車内、オフィススペース内及び歩道上といった環境とした。評価方法は、①無音声区間フレームを音声区間フレームであると誤判別した場合、②発話区間の始端・終端において、発話区間を無発話区間であると誤判別した場合、このような①や②に該当する場合のフレームをエラーフレームとした。また、比較対象 (従来例) として、平均ゼロ交差回数と対数パワーとを変数としたフィッシャーの線形判別関数による手法を用いた。

- 25      図 7 は、その結果を示す。この図 7 は、総フレームに対するエラーフレームの割合の百分率 (発話区間誤検出率) を示す。図 7 中、LDF の値は、前記線形判別関数による手法の値であり、CRS の値はクロススペクトルを用いた手法 (本発明) の値である。

この図 7 に示すように、防音室内やオフィススペース内においては、発話区間誤検出率の結果に、平均ゼロ交差回数と対数パワーによる方法と本発明による手法とで大きな差はみられない。しかし、自動車内や歩道では、発話区間誤検出率

の結果が本発明による手法により改善される結果を示すようになった。このように、本発明は、特に雑音環境下において有効に作用する。

次に第2の実施の形態を説明する。

図8は、この第2の実施の形態の音声信号処理装置10の構成を示す。この第2の実施の形態では、第1マイク1と第2マイク2とで受音した音声信号を合成して後段の音声アプリケーションに出力する構成になっている。このため、この第2の実施の形態では、遅延処理部51と波形合成部52とを備え、遅延処理部51で第2マイク2からの音声信号を遅延させて波形合成部52に出力して、波形合成部52で、遅延処理部51で遅延されて入力された第2マイク2の音声信号と第1マイク1からの音声信号とを合成して出力している。

第1マイク1と第2マイク2といった複数のマイクで受音した音声信号間には、音源から各マイク1, 2までの距離の違いに起因する位相差が生じる。このようなことから、第1マイク1と第2マイク2といった複数のマイクで受音した音声信号を合成しようとする場合には、音源から各マイク1, 2までの音声信号の到達時間差を補正し、位相を同相化したのちに音声信号を加算する、という遅延処理が必要になる。このようなことから、前述したように、第2の実施の形態では、遅延処理部51と波形合成部52とを備えている。

そして、前述の第1の実施の形態では（図6参照）では、最頻値計算部33Cがヒストグラムから最頻傾き $\tau_0$ を算出しているが、第2の実施の形態では、そのような最頻傾き $\tau_0$ に基づいて、遅延処理部51で遅延処理しているのである。以下に具体的に説明する。

前記図3や図4中（A）に示すように音声区間ではクロススペクトルの位相成分が一定の傾きを有するが、この傾きは、第1マイク1と第2マイク2とのチャネル間の遅延時間を示すものとなる。

このような関係を利用して、遅延処理部51では、ヒストグラム等計算部33が算出した前記最頻傾き $\tau_0$ に基づいて、遅延処理している。具体的には、図6に示すように、最頻値計算部33Cから遅延処理部51に最頻傾き $\tau_0$ が出力されており、遅延処理部51は、入力されたこの最頻傾き $\tau_0$ に基づいて遅延処理している。

$$\tau_0 = x / n = 2\pi \cdot n_0 / N \quad [\text{rad/point}] \quad \dots (5)$$

ここで、 $x$ 、 $n$ の単位はそれぞれラジアン、周波数ポイント (point) であり、 $N$ は、FFTポイント数であり、 $n_0$ は遅延サンプリングポイント数である。

この関係から、下記 (6) 式として、最頻傾き  $\tau_0$  を変数とした遅延サンプリングポイント数  $n_0$  を得ることができる。

$$n_0 = \tau_0 / (2\pi / N) \quad [\text{point}] \quad \dots (6)$$

そして、この遅延サンプリングポイント数  $n_0$  を用いて、下記 (7) 式により、遅延時間  $t_0$  を得ることができる。

$$t_0 = n_0 / F_s \quad \dots (7)$$

10      ここで、 $F_s$  は、サンプリング周波数であり、例えば 16 kHz である。

遅延処理部 51 では、このようにして得た遅延時間  $t_0$  に基づいて、入力される第 2 マイク 2 の音声信号を遅延して、波形合成部 52 に出力する。

波形合成部 52 は、遅延処理部 51 で遅延されて入力された第 2 マイク 2 の音声信号と第 1 マイク 1 からの音声信号とを合成して出力する。

15      なお、音声信号の合成信号を次のようにして得ることもできる。

前述したように、周波数関数  $X_2$  に遅延項  $e^{j\omega t_0}$  をかけた  $X_2(\omega) e^{j\omega t_0}$  は、周波数関数  $X_1$  と同相化され、これにより、 $X_1(\omega) + X_2(\omega) e^{j\omega t_0}$  の逆フーリエ変換をチャネル同期加算音声として扱うことができる。この関係を利用して、音声信号の合成信号を得る。

20      すなわち、先ず遅延時間  $t_0$  を用いることで、下記 (8) 式により、周波数軸上でチャネル同期加算音声  $X_1(\omega) + X_2(\omega) e^{j\omega t_0}$  を得る。ここで、遅延時間  $t_0$  は、前記 (6) 式及び (7) 式に示すように最頻傾き  $\tau_0$  を変数とする値である。

$$\begin{aligned} X_1(\omega) + X_2(\omega) e^{j\omega t_0} = & \{ \text{Re}[X_1(\omega)] + j \text{Im}[X_1(\omega)] \} + \{ \text{Re} \\ 25 & [X_2(\omega)] (\cos \omega t_0 + j \sin \omega t_0) + j \text{Im}[X_2(\omega)] (\cos \omega t_0 + j \sin \omega t_0) \} \quad \dots (8) \end{aligned}$$

ここで、チャネル同期音声スペクトルは、実部、虚部にそれぞれ

$$\begin{aligned} \text{Re} : & \text{Re}[X_2(\omega)] \cos \omega t_0 - \text{Im}[X_2(\omega)] \sin \omega t_0 + \text{Re}[X_1(\omega)] \\ \text{Im} : & \text{Re}[X_2(\omega)] \sin \omega t_0 + \text{Im}[X_2(\omega)] \cos \omega t_0 + \text{Re}[X_1(\omega)] \end{aligned}$$

を持つ複素スペクトルになる。この処理をフレーム毎に施し、それぞれのフレーム毎に IFFT (インバースFFT) をし、同期加算音声のフレーム列を得る。

そして、そのようにして得たフレーム列にオーバーラップアッド法 (Overlap-add method) を適用して同期加算音声、すなわち第 1 マイク 1 の音声信号と第 5 2 マイク 2 の音声信号との合成信号を得る。

ここで、オーバーラップアッド法とは、図 9 に示すように、入力データ列  $s_n(t)$  を重ね合わせながら加算する方法である。ここで、 $s_n(t)$  は  $n$  番目の合成音声波形フレームを示す。また、図中  $L$  は定数である。

10 以上のように音声信号処理装置 10 を構成することで、遅延処理部 51 が第 2 マイク 2 からの音声信号を遅延させて波形合成部 52 に出力して、波形合成部 52 が、遅延処理部 51 により遅延されて入力された第 2 マイク 2 からの音声信号と第 1 マイク 1 からの音声信号とを合成して出力する。

これによる効果は次のようになる。

15 前述の第 1 の実施の形態で説明したように、クロススペクトルの位相の周波数に対する傾きは、音源と第 1 マイク 1 までの距離と音源と第 2 マイク 2 までの距離との差に対応して変化する値である。このようなクロススペクトルの位相の周波数に対する傾きから前記遅延時間を推定している。そして、実際に推定の際に用いる値を、最頻傾き  $\tau_0$  としている。このように最頻傾き  $\tau_0$  を用いて、遅延時間を推定しているので、精度を高くして遅延時間の推定を行うことができる。

20 そして、このような遅延時間に基づいて、第 1 マイクと第 2 マイクとの音声信号を合成することで、高品質の合成音声信号を提供することができる。例えば、このような合成音声信号を利用した場合、音声認識システムでは、高認識率、低誤認識率の音声認識が可能になり、また、携帯電話やトランシーバでは、高品質の音声による通話が可能になり、放送システムでは、高品質の放送や録音が可能  
25 になる。

また、遅延時間を推定に用いる前記傾きを、最頻傾き  $\tau_0$  とした結果、前述の第 1 の実施の形態と同様に、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現すること

ができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

- 5 以上のように、小型・軽量で脱着が容易であり、接話マイクとほぼ同等の近距離音声を確認することができ、接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる装着型マイクを用いることを前提としつつも、前述した種々の効果を得ることができる。

(実施例 (第2の実施の形態))

本発明を適用したシステムにより生成した同期加算音声 (合成音声信号) を用いて、音響モデルによる音声認識の実験をした。

- 10 音響モデルによる音声認識実験では、先ず、同期加算音声による学習データにより、音響モデルを作成した。作成した音響モデルは次のようになる。

①収録環境毎に作成した4種類の収録環境依存型HMM (hidden Markov model)

②すべて環境の収録音声により学習した収録環境非依存型HMM

- 15 ここで、前記収録環境とは、前記防音室内、自動車内、オフィススペース内及び歩道上である。

そして、作成した音響モデルを用いて、音声認識実験を行った。

認識タスクは連続音声認識であり、評価用データ (評価用音声) は学習時と異なる音声としている。図10は、その音声認識実験で得た認識率の結果を示す。

- 20 ここで、比較対象 (従来例) として、第1マイク1と第2マイク2とからの単チャンネル音声による認識率の結果も示す。例えば、第1マイク1は眼鏡マイクであり、第2マイク2は胸元マイクである。ここで、眼鏡マイクとは、眼鏡のフレームに装着したマイクである。

- 25 この図10に示すように、車内以外の、防音室内、歩道上及びすべての環境で、本発明により得た同期加算音声による認識率が、単チャンネル音声の認識率を上回る結果となっている。これにより、実環境においても、本発明を適用したシステムが生成した同期加算音声が高品質であることがわかる。

次に第3の実施の形態を説明する。

図 11 は、この第 3 の実施の形態の音声信号処理装置 10 の構成を示す。この第 2 の実施の形態の音声信号処理装置 10 は、前述の第 1 の実施の形態の音声信号処理装置 10 の構成と、第 2 の実施の形態の音声信号処理装置 10 の構成とを組み合わせた構成になっている。すなわち、第 3 の実施の形態の音声信号処理装置 10 は、音声／非音声判定部 34、遅延処理部 51、波形合成部 52 及び音声入力オン／オフ制御部 18 を同時に備えている。

このように構成することで、第 3 の実施の形態の音声信号処理装置 10 は以下のように動作する。なお、特に言及しない部分については、前述の第 1 の実施の形態の音声信号処理装置 10 や第 2 の実施の形態の音声信号処理装置 10 と同様に動作するものとする。

遅延処理部 51 が、ヒストグラム等計算部 33（最頻値計算部 33C）が算出した最頻傾き  $\tau_0$  に基づいて、第 2 マイク 2 の音声信号を遅延し、波形合成部 52 が、遅延処理部 51 で遅延されて入力された第 2 マイク 2 からの音声信号と第 1 マイク 1 からの音声信号とを合成して、合成音声信号を音声入力オン／オフ制御部 18 に出力する。

一方、音声／非音声判定部 34 が、ヒストグラム等計算部 33 が得た頻度に基づいて、音声区間と非音声区間とを判定し、音声入力オン／オフ制御部 18 では、その判定結果に基づいて、波形合成部 52 から出力される音声信号（同期加算音声信号）の出力をオン又はオフする。

このように構成することで、第 3 の実施の形態の音声信号処理装置 10 は、前述の第 1 の実施の形態の音声信号処理装置 10 が有する効果と、第 2 の実施の形態の音声信号処理装置 10 が有する効果とを発揮することができる。

すなわち、高品質の合成音声信号を生成するとともに、その合成音声信号中の発話区間を精度よく検出することができる。さらに、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

以上、本発明の実施の形態について説明した。しかし、本発明は、前述の実施の形態として実現されることに限定されるものではない。

例えば、図 1 2 に示すように、前記音声／非音声判定部 3 4 が、第 1 乃至第 N 傾き計算部 3 2<sub>1</sub>〜3 2<sub>N</sub> が算出した前記傾き  $\tau_i$  と最頻傾き  $\tau_0$  とを下記 (9) 式により比較する。

$$|\tau_i - \tau_0| < \alpha \sigma \quad \dots (9)$$

- 5      ここで、 $\alpha$  は係数であり、 $\sigma$  は前記判定用の閾値（傾き閾値） $\delta$  に物理的に内包される値である。例えば、 $\delta$  と  $\alpha \sigma$  とを用意した意味は、 $\delta$  を固定値とし、 $\alpha \sigma$  をリアルタイム学習により随時更新する変数とし、これにより、各値による音声区間の検出の効果の違いを区別するためである。

- 10       $\alpha \sigma$  の  $\sigma$  を更新することで、静粛な環境では、音声区間判定条件を厳しくし、より非音声区間の誤判定を防止することができる。すなわち、バックグラウンドノイズのある環境では判定条件を甘くすることで、音声区間を安定して検出することが可能になる。仮に、バックグラウンドノイズのある環境にもかかわらず静粛環境の  $\sigma$  を用いてしまうと、この場合固定値の  $\delta$  を用いることと等価となるが、この場合には、ノイズと音声とが重なっているようなとき、音声区間が棄却されて  
15      しまうおそれがある。

すなわち、固定値としての  $\delta$  は、その値を設定した条件に近い環境での音声区間を検出に用いるときに当該音声区間の検出に有効に作用し、変数である  $\alpha \sigma$  は、環境の変化に対し動的に対応するシステムに用いるときに音声区間の検出に有効に作用する。

- 20      また、係数  $\alpha$  を変更することでも、判定を厳しくしたり、甘くしたりすることもできる。

また、前述の実施の形態では、前記帯域毎の傾きをヒストグラム化することで、帯域毎のそれぞれの傾きが特定の傾きに集中する傾向をみている。しかし、他の手法により、帯域毎のそれぞれの傾きが特定の傾きに集中する傾向をみるよう  
25      にしてもよい。

また、前述の実施形態では、検出対象音が人間が発する発話音である場合を説明したが、検出対象音は、人間以外の物体が発する音でもよい。

また、前述の実施の形態の説明において、第 1 及び第 2 フレーム化部 1 1, 1 2、第 1 及び第 2 周波数分析部 1 3, 1 4 及びクロススペクトル計算部 1 5 が、

複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段を実現しており、位相抽出部16、位相unwrap処理部17、周波数帯域分割部31及び第1乃至第N傾き計算部32<sub>1</sub>〜32<sub>N</sub>が、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段を実現しており、ヒストグラム等計算部33及び音声／非音声判定部34が、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した発話音の発話区間を検出する発話音検出手段を実現している。

また、ヒストグラム等計算部33及び遅延処理部51が、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での受音の遅延時間を検出する遅延時間検出手段を実現しており、波形合成部52が、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段を実現している。

また、前述の実施形態の音声信号処理装置10を音声認識装置に適用することができる。この場合、音声認識装置は、前述したような音声信号処理装置10の構成に加えて、音声信号処理装置10が検出した発話区間の音声信号（発話音）について音声認識処理をする音声認識処理手段を備える。

ここで、音声認識技術としては、例えば、旭化成株式会社が提供する音声認識技術「VORERO」（商標）（<http://www.asahi-kasei.co.jp/vorero/jp/vorero/feature.html>参照）等があり、このような音声認識技術の用いた音声認識装置に適用することもできる。

また、前述の実施形態の音声信号処理装置10をコンピュータで実現することができる。そして、前述したような音声信号処理装置10の処理内容をコンピュータが所定のプログラムにより実現する。この場合、プログラムは、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検



- 出対象音を検出する処理をコンピュータに実行させるプログラムになる。又は、プログラムは、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生
- 5 する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させるプログラムになる。

#### 産業上の利用の可能性

- 10 本発明によれば、装着型マイクロホンを用いた環境変動に対してもロバストな受音系の構築を可能にすることができる。

## 請求の範囲

1. 検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音を検出することを特徴とする対象音検出方法。
- 5 2. 前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記検出対象音を検出することを特徴とする請求の範囲第1項記載の対象音検出方法。
- 10 3. 前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに検出対象音を検出することを特徴とする請求の範囲第2項記載の対象音検出方法。
- 15 4. 複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする請求の範囲第1乃至3項のいずれかに記載の対象音検出方法。
- 20 5. 音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出することを特徴とする信号入力遅延時間検出方法。
- 25 6. 前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記受音の遅延時間を検出することを特徴とする請求の範囲第5項記載の信号入力遅延時間検出方法。
7. 前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに、前記受音の遅延時間を検出することを特徴とする請求の範囲第6項記載の信号入力遅延時間検出方法。

8. 複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする請求の範囲第5乃至7項のいずれかに記載の信号入力遅延時間検出方法。
9. 複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、
- 5 前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、
- 前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した検出対象音源から出力された検出対象音を検出する対象
- 10 音検出手段と、
- を備えたことを特徴とする音信号処理装置。
10. 前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、
- 前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づ
- 15 いて、前記検出対象音を検出することを特徴とする請求の範囲第9項記載の音信号処理装置。
11. 音源から出力された音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された音进行处理する音信号処理装置において、
- 前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検
- 20 出するクロススペクトル位相検出手段と、
- 前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、
- 前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する遅延時間検出手段
- 25 と、
- 前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、
- を備えたことを特徴とする音信号処理装置。
12. 前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割

した帯域毎に傾きを検出しており、

前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出することを特徴とする請求の範囲第 11 項記載の音信号処理装置。

- 5    13.    検出対象音源から出力された検出対象音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された検出対象音を処理する音信号処理装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、

- 10    前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記検出対象音源からの受音の遅延時間を検出する遅延時間検出手段と、

- 15    前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記検出対象音を検出する対象音検出手段と、

- 20    を備えたことを特徴とする音信号処理装置。

14.    前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前

- 25    記検出対象音を検出することを特徴とする請求の範囲第 13 項記載の音声信号処理装置。

15.    発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音を処理する音声認識装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検

- 出するクロススペクトル位相検出手段と、  
前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、  
前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数の
- 5    マイクロホンが受音した前記発話音を検出する発話音検出手段と、  
前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、  
を備えたことを特徴とする音声認識装置。
16.    前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、  
前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする請求の範囲第15項記載の音声認識装置。
17.    発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音処理する音声認識装置において、  
前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、  
前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、
- 20    前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記発話源からの受音の遅延時間検出する遅延時間検出手段と、  
前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、
- 25    前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記発話音を検出する発話音検出手段と、  
前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、  
を備えたことを特徴とする音声認識装置。

18. 前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする請求の範囲第17項記載の音声認識装置。
19. 検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検出対象音を検出する処理をコンピュータに実行させることを特徴とするプログラム。
20. 音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させることを特徴とするプログラム。

図 1

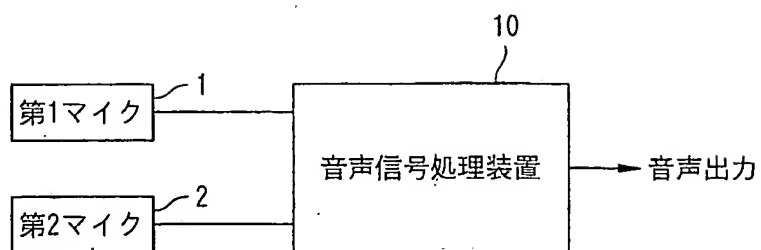


図 2

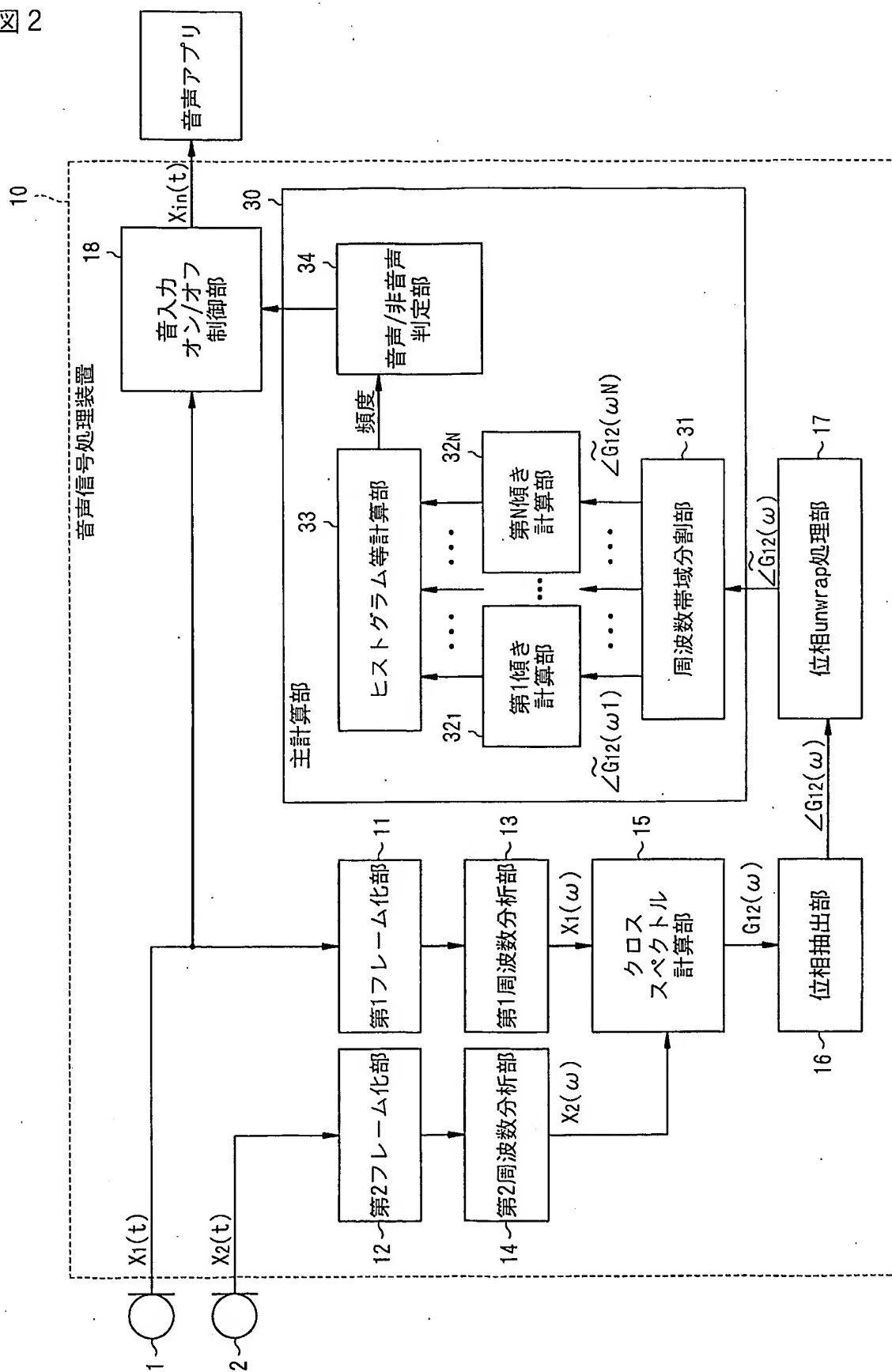




図 3

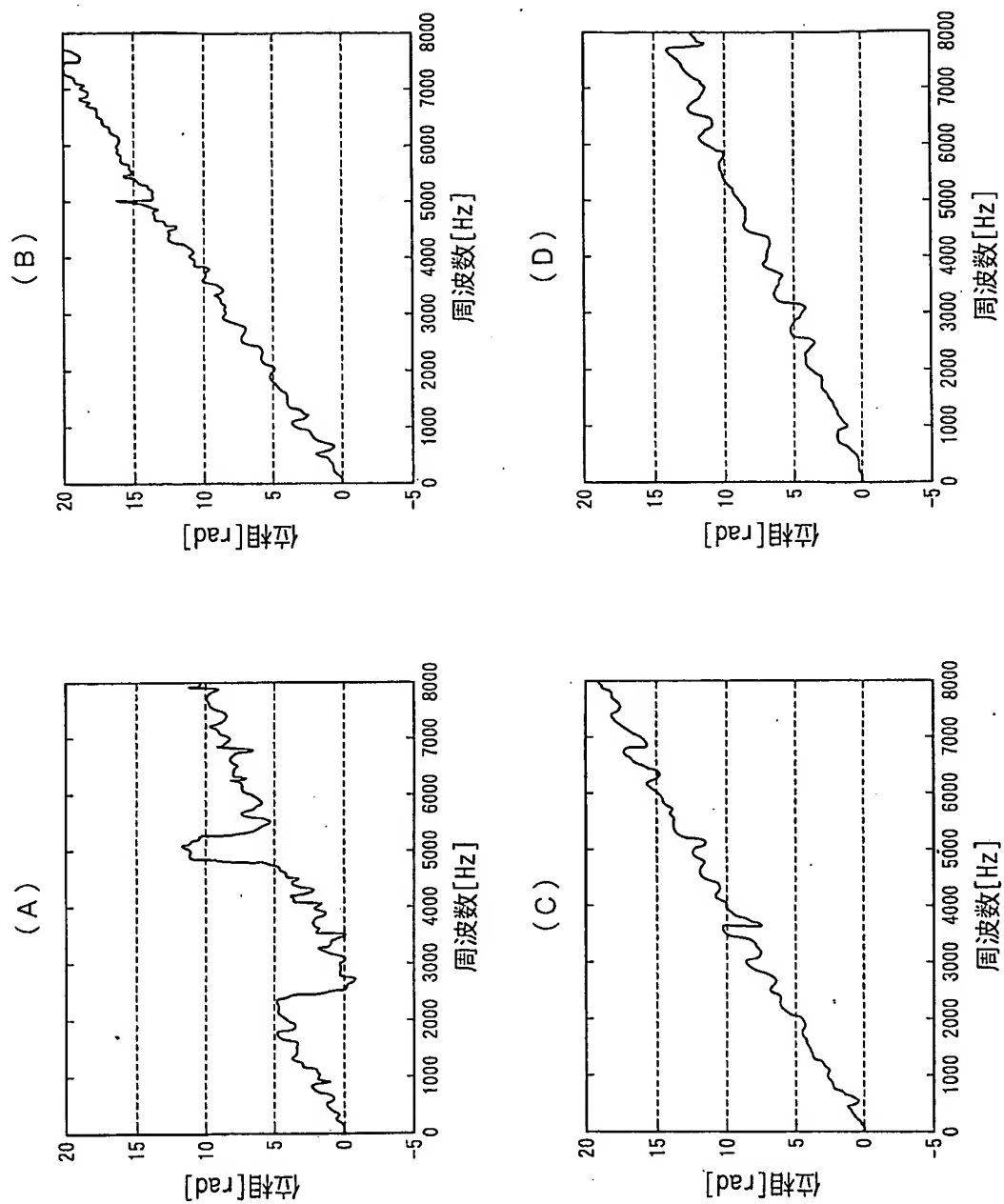


図 4

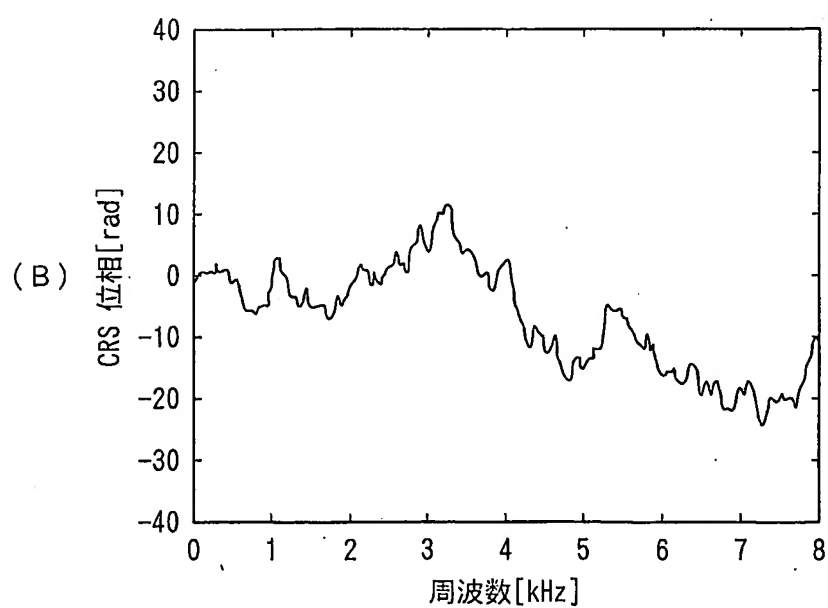
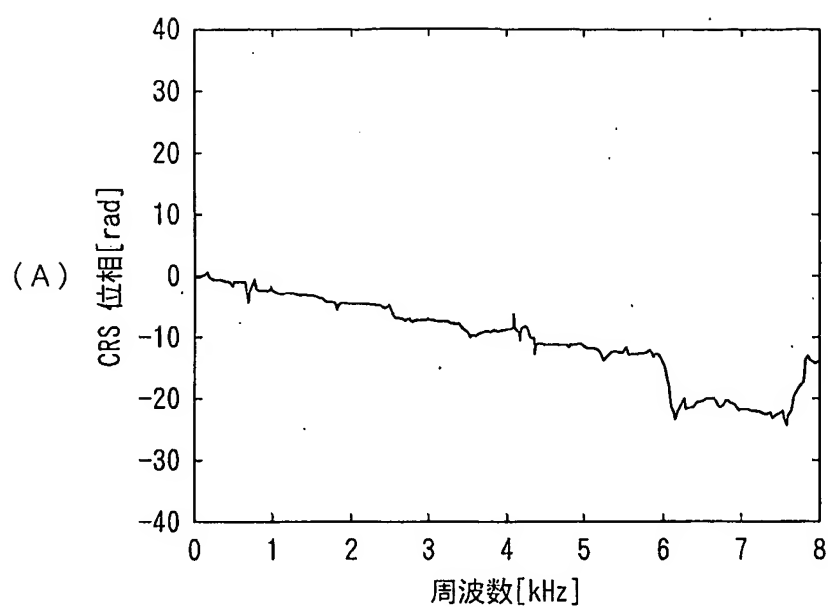


図 5

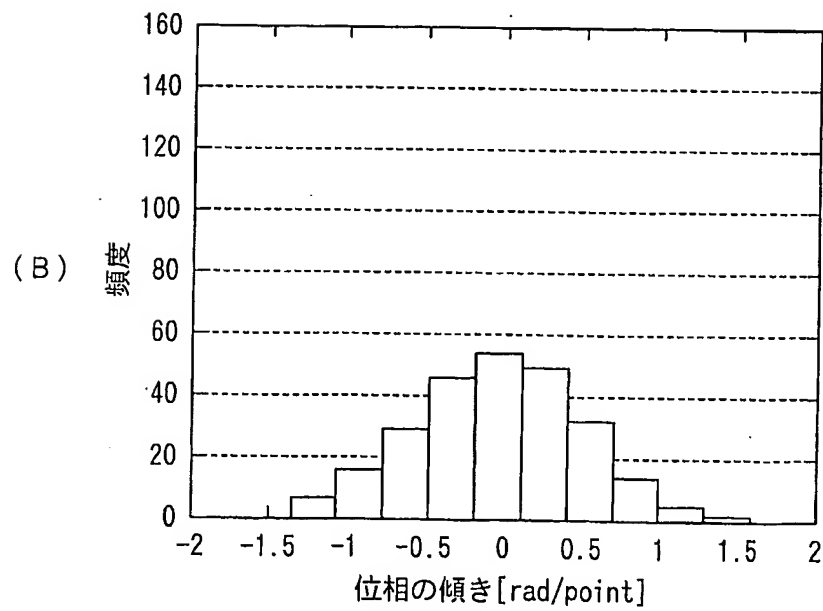
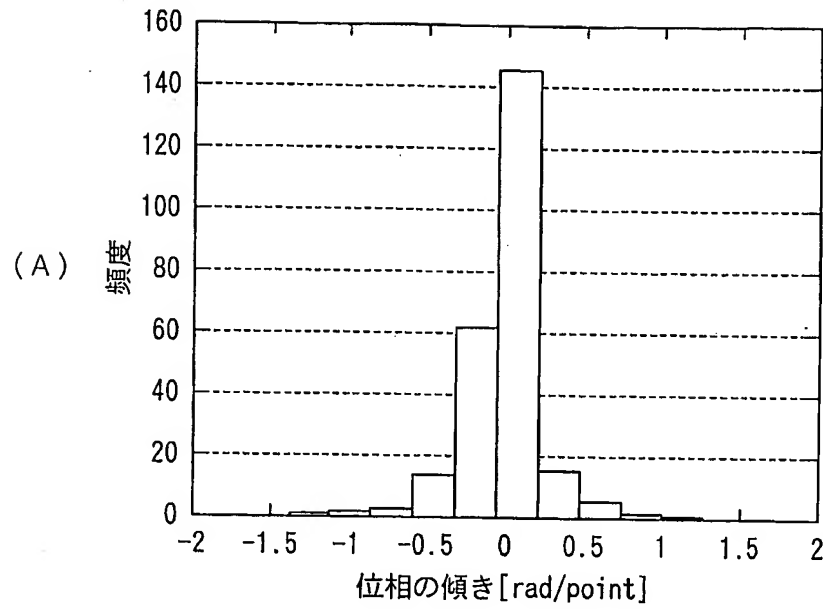


図 6

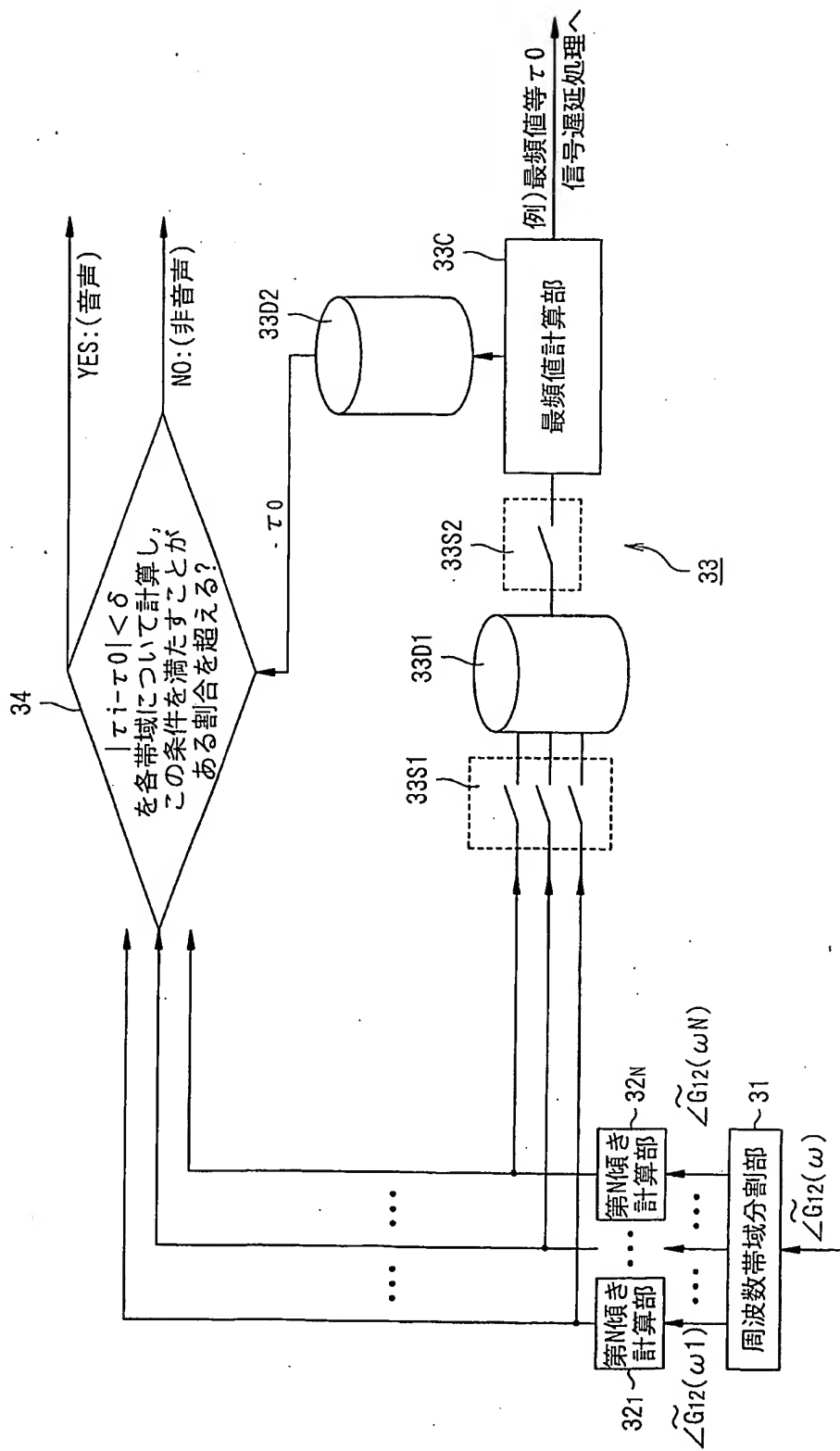


図 7

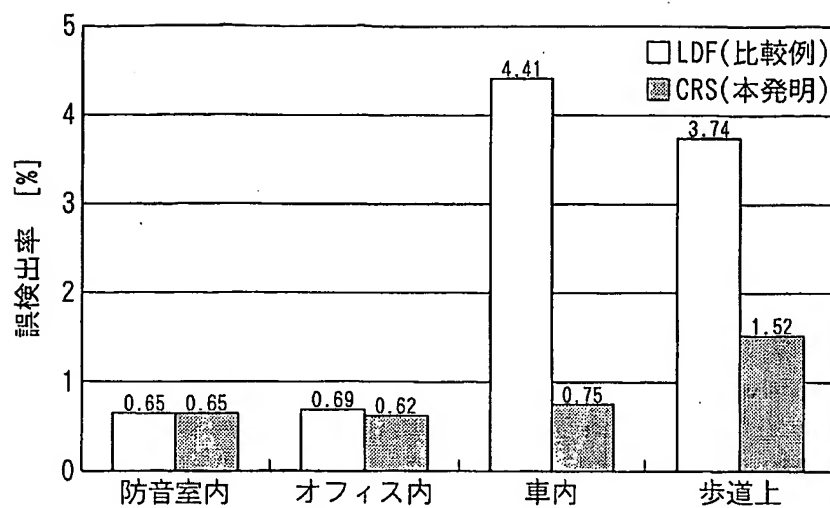


図 8

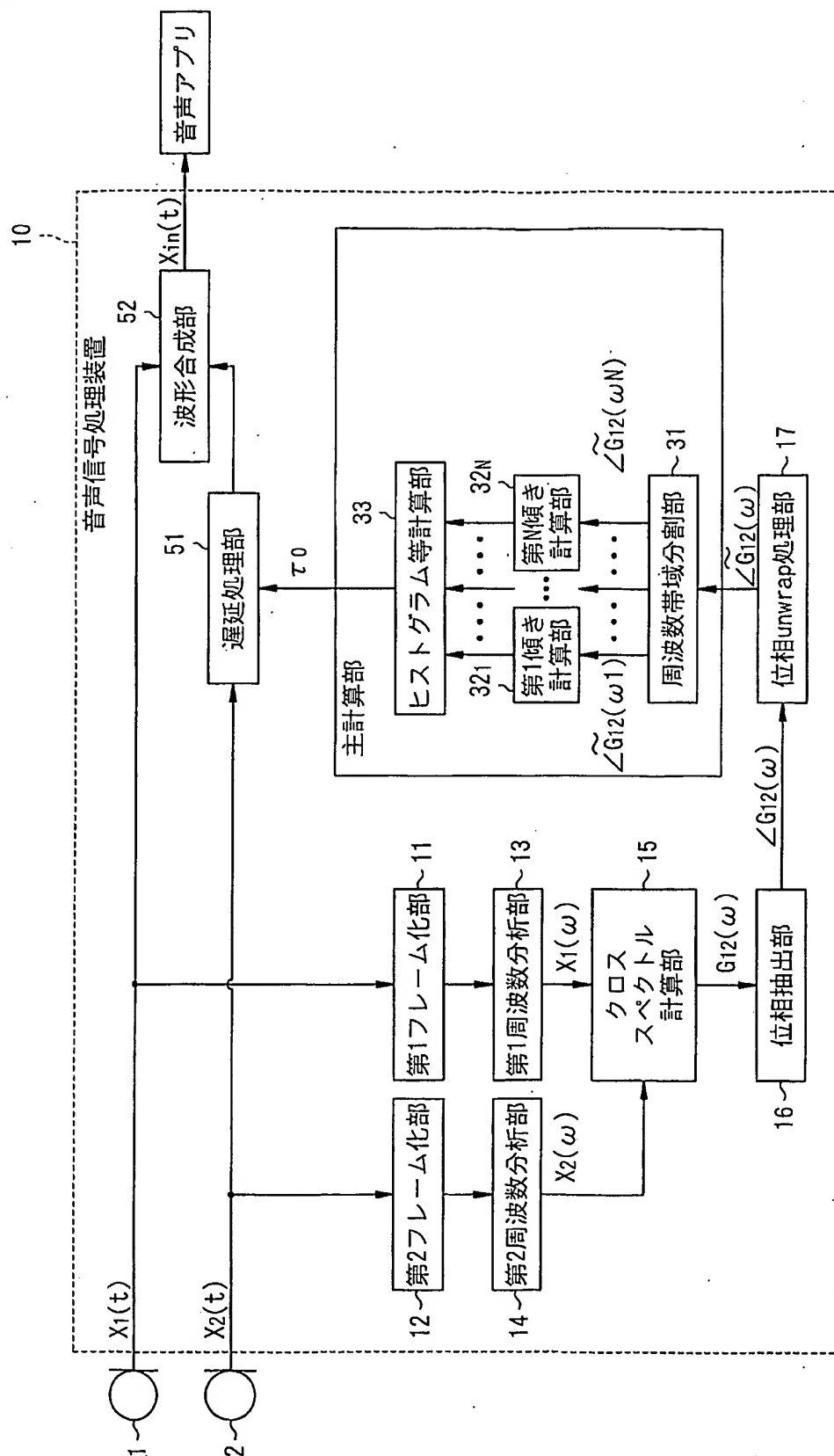


図 9

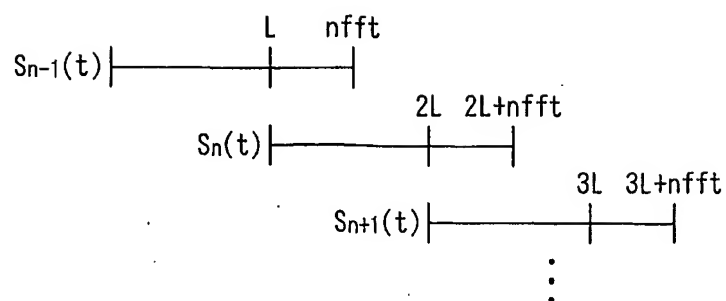


図 10

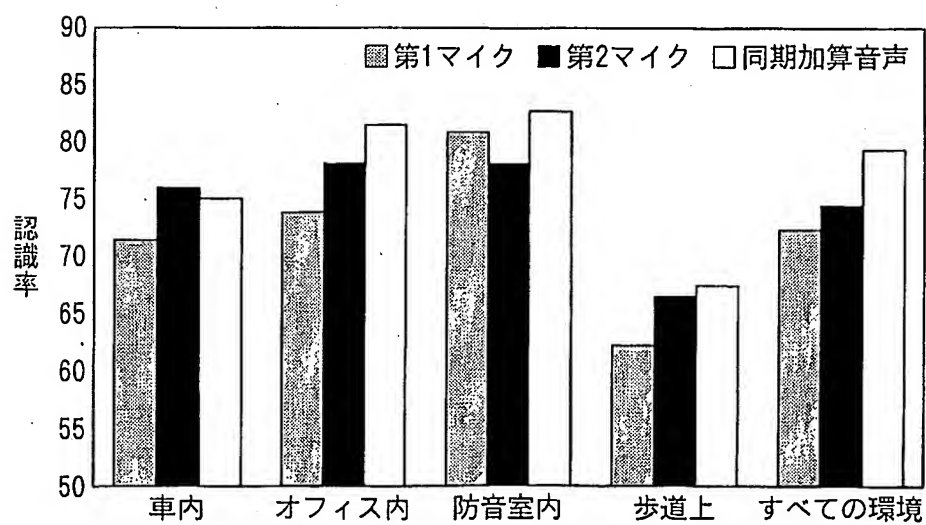


図 1 1

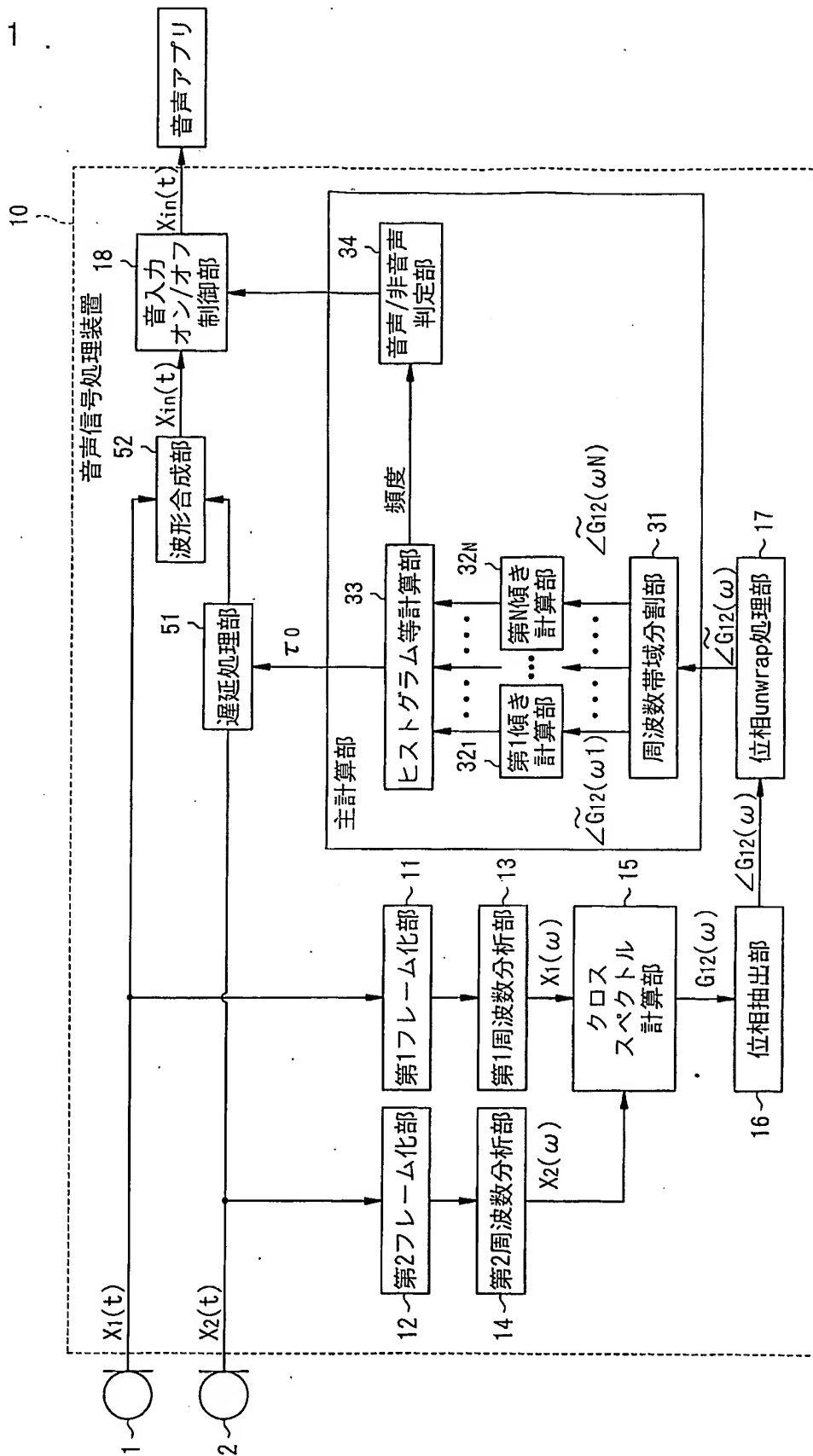
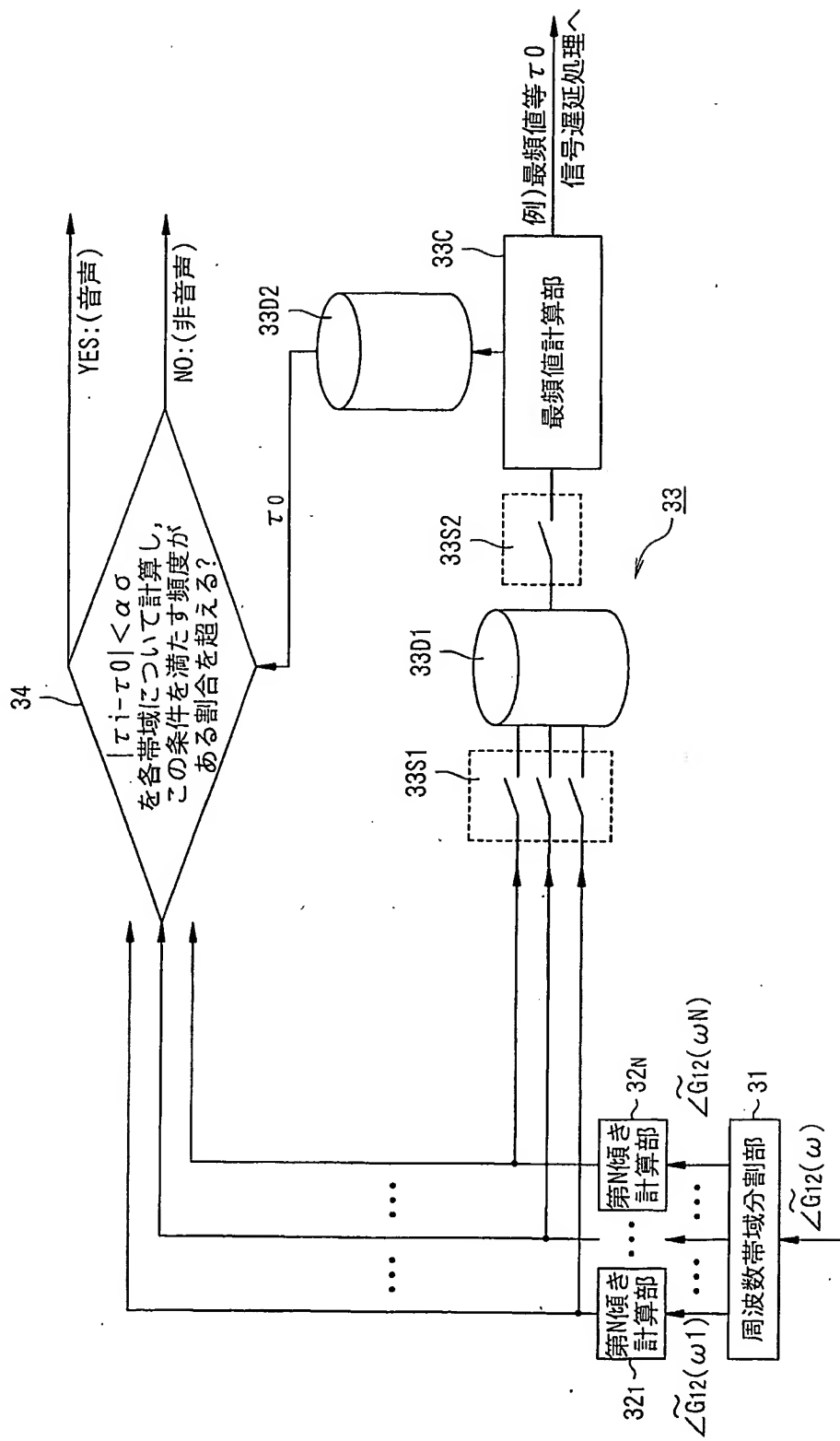




図 1 2



## INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/003524

## A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl<sup>7</sup> G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int.Cl<sup>7</sup> G10L21/02, G01N29/18

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Toroku Jitsuyo Shinan Koho	1994-2004
Kokai Jitsuyo Shinan Koho	1971-2004	Jitsuyo Shinan Toroku Koho	1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

JSTPlus FILE (JOIS), IEEE Xplore

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	M. Omologo, P.Svaizer, 'Acoustic event localization using a crosspower-spectrum phase based technique', Proc.of ICASS94, 1994, pages II-273 to II-276	1, 2, 4-6, 8-12, 15, 16, 19, 20
Y	JP 9-196900 A (Hitachi, Ltd.), 31 July, 1997 (31.07.97), (Family: none)	1, 2, 4-6, 8-12, 15, 16, 19, 20
P, X	Kiyoshi TATARA et al., "Fukusu no Sochaku Microphone o Mochiita Juonkei no Kochiku ni Kansuru Kento", The Acoustical Society of Japan (ASJ) 2003 nen Shunki Kenkyu Happyokai Koen Ronbunshu I, 18 March, 2003 (18.03.03), 3-Q-13, pages 117 to 178	1-20

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search  
06 May, 2004 (06.05.04)Date of mailing of the international search report  
25 May, 2004 (25.05.04)Name and mailing address of the ISA/  
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

A. 発明の属する分野の分類 (国際特許分類 (IPC))		
Int. Cl <sup>7</sup> G10L21/02		
B. 調査を行った分野		
調査を行った最小限資料 (国際特許分類 (IPC))		
Int. Cl <sup>7</sup> G10L21/02		
Int. Cl <sup>7</sup> G01N29/18		
最小限資料以外の資料で調査を行った分野に含まれるもの		
日本国実用新案公報 1922-1996年		
日本国公開実用新案公報 1971-2004年		
日本国登録実用新案公報 1994-2004年		
日本国実用新案登録公報 1996-2004年		
国際調査で利用した電子データベース (データベースの名称、調査に使用した用語)		
JSTPlusファイル (JOIS)		
IEEE Explore		
C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	M. Omologo, P. Svaizer, 'Acoustic event localization using a crosspower-spectrum phase based technique', Proc. of ICASSP94, 1994, p. II-273~II-276	1, 2, 4-6, 8-12, 15, 16, 19, 20
Y	JP 9-196900 A (株式会社日立製作所) 1997. 7. 31 (ファミリーなし)	1, 2, 4-6, 8-12, 15, 16, 19, 20
<input checked="" type="checkbox"/> C欄の続きにも文献が列挙されている。 <input type="checkbox"/> パテントファミリーに関する別紙を参照。		
* 引用文献のカテゴリー 「A」 特に関連のある文献ではなく、一般的技術水準を示すもの 「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの 「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す) 「O」 口頭による開示、使用、展示等に言及する文献 「P」 国際出願日前で、かつ優先権の主張の基礎となる出願		
の日の後に公表された文献 「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの 「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの 「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの 「&」 同一パテントファミリー文献		
国際調査を完了した日	国際調査報告の発送日	
06. 05. 2004	25. 5. 2004	
国際調査機関の名称及びあて先	特許庁審査官 (権限のある職員)	5 C 3 1 4 8
日本国特許庁 (ISA/JP)	渡邊 聡	
郵便番号100-8915	電話番号 . 03-3581-1101	内線 3541
東京都千代田区霞が関三丁目4番3号		

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
PX	<p>多々良潔, 他, 複数の装着型マイクロホンを用いた受音系の構築に関する検討, 日本音響学会2003年春季研究発表会講演論文集I, 2003. 03. 18, 3-Q-13, p. 177-178</p>	1-20